Discovering Correlations Across Data Sets
Final Report

Hiroshi Horikawa
Spatial and Temporal
hiroshi.horikawa@nyu.edu

Zipei Bian
Spatial and Temporal
zb534@nyu.edu

Pyay Aung San
Spatial and Temporal
pas502@nyu.edu

ABSTRACT

Every data sets contain information such as date, location, money involved, speed, precipitation, luminosity, etc. The data collected by public and private institutions contain mass amount of information, but they only provide information that is directly related to the data that was collected. While these can be useful to understand what was happening when the data was collected, it is impossible to understand if there are any external factors that was affecting the data. For example, a data set containing the weather information can tell the precipitation, temperature, and wind speed of a day while a data set containing the crime information can tell what crime happened where on which date and time. The weather data is unable to explain anything about crime while the crime data is unable to explain the precise meteorological condition when the crime occurred. But using these two data sets, we may discover new set of information such as understanding under how wind speed, precipitation, or temperature affect crime.

1 INTRODUCTION

There are many things that has correlation with one another. There have been studies which used correlation across data sets to understand and determine the root of a problem. For example, there has been a research to understand features of crime rates using big data. Given that this paper is for a course which has emphasis on applying big data to solve real world problems, we are using multiple data sets which must be analyzed through a cluster of computers.

For this paper, we are presenting correlations across different data sets. Some of the applications we used to clean data are OpenRefine and Google Dataprep. The platform we used to aggregate the data is Dumbo where we used Spark and Hadoop. Finally, the correlations across the aggregated data sets are determined using IPython Notebook which can implement libraries such as Numpy and Pandas to make the analysis.

For the methodology to find correlation across data sets, we will use Spearman's rank order correlation (simplified as Spearman's correlation later). Instead of writing out the entire equation, we will use a method from the Pandas library which computes Spearman's correlation. The output of the method will be the correlation coefficient and the p-value to test for non-correlation. Both the correlation coefficient and p-value will be returned as a float value.

1.1 Problem Formulation

For this paper, we are trying to understand how data sets correlate with one another. More specifically, we will center on the NYPD Complaint Data Historic (which we will refer to as crime data set) and discover the factors that correlate with the crime data. We will be computing correlations across different data sets (such as data sets for weather, vehicle collision, property price, etc.) to see if there is anything that affects where and when crimes occur (hence our group name, "Spatial and Temporal"). As noted previously, we will use Spearman's correlation to compute the correlation.

1.1.1. NYPD Complaint Data Historic. Our group chose to center our research on the crime data set since public safety is important to the lives of people and reputation of the city. But the city government is not able to mobilize large numbers of law enforcements to every neighbor in the city due to budget constraints. It is important to know when and where to focus the number of officers to enforce law efficiently as it will both increase the level of safety per dollar spent.

2 EXPERIMENTAL AND COMPUTATIONAL DETAILS

2.1 Mathematical Aspects

As mentioned previously, we have used Spearman's correlation to find the correlations across data sets. This correlation method is like Pearson's correlation coefficient but computes the correlation coefficient between two ranked variables. In other words, Spearman's correlation is the Pearson's correlation for the ranks of the variables. It means that Spearman's correlation doesn't need the assumption that the variables have linear relationship and the variables doesn't have to be on interval scales.

We did not normalize the variables on the data set since we were interested in the monotonic relations between two variables (also, Spearman's correlation doesn't require normalization). The coefficient is bounded between –1 and 1 where –1 means inverse correlation, 1 means high correlation, and 0 means no correlation between the data sets. [Hauke 2011]

Spearman's rank-order correlation coefficient is defined as the following equations:

$$r_s = \rho_{rg_X, rg_Y} = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}}$$

- r and ρ denotes the Pearson's correlation coefficient but is utilized to the rank variables.
- $cov(rg_X, rg_Y)$ is the covariance of the rank variables
- σ_{rgX} and σ_{rgX} denotes the standard deviations of the rank variables

In the case where all n ranks are distinct integers, it can be calculated using the formula

$$r_s = 1 - \frac{6\Sigma d_i^2}{n(n^2 - 1)}$$

- $d_i = rg(X_i) - rg(Y_i)$ is the difference of each ranks at each observation
- *n* is the number of observations

2.2 Methodology, architecture, and design

2.2.1 Data Cleaning using tools. The data must be cleaned prior to the analysis. We have used applications such as OpenRefine and Google Dataprep to take care of any null values and clustered values that should be consistent with one another. We use OpenRefine to clean up small-scale data such as the one for weather. However, to handle larger data set such as crime, we use a beta release Google Dataprep which is similar to OpenRefine. Google Dataprep also provides a comprehensive descriptive statistic to the summary of the whole data sets. The application is managed by Trifacta, a third-party data wrangling software, which is embedded in Dataprep. We will provide a quick How-to guide to use Dataprep in a separate document in the final submission.

2.2.2 Descriptive trends using Tableau. After understanding the quality of the data, we use Tableau to quickly graph the trends of the data sets. The advantage of using Tableau BI tool is providing us descriptive stats efficiently with filters so that we can understand the rough nature of the data sets, without coding intensively. However, we do use PySpark and IPython notebook to find the correlation across each data sets.

2.2.3 Data Preparation using PySpark and Hadoop. After getting the descriptive trends using Tableau, we used Python with either PySpark or Hadoop and prepared the aggregated output of each data set (such as the number of crime by year) before we find the correlation across other data sets. We host original data sets on Dumbo, do the operation and download the necessary output file to local computer to do aggregated analysis. The codes on how we prepare data sets can be found in GitHub under Dataprep folder.

2.2.4 Query Tool using Google BigQuery. In addition to preparing data in PySpark and Hadoop, we found BigQuery a simple and efficient tool to obtain aggregated outputs. We uploaded or transferred the data sets to the Google Cloud Storage (GSC). We then created tables in BigQuery by loading the data sets stored in GSC. For NYC Yellow Taxi, which had already been published as a merged data set, we simply connected it with our BigQuery without merging all the monthly data sets and uploading over 150GB data which would be taxing and exceeding the free tier limit.

3 ISSUES WITH DATA

3.1 Data issues

3.1.1 NYPD Complaint Data Historic data issues. Dataprep gives us a descriptive quality summary of each column. For instance, at CMPLNT_FR_DT (column 1) column, plotted in Figure 1, we will see that 655 rows are missing, 29 rows are mismatched and date ranges from 1900 to 2016. Mismatched values in Dataprep represent any rows which do not follow the majority (in this case date format). This column represents the actual, not necessarily reported, date when the incident occurred. Even though some incidents happened in a single date, some occurred within a period and end time records are in CMPLT_TO_DT (column 3). RPT_DT (at column 5) gives the single dates when police recorded the occurrence. Since there is no missing nor mismatched values for column 5, we use this column to have our preliminary experimentation to find the correlation with other data sets such as weather or collisions.

However, we cannot stop there. We need to examine whether the dates from column 1 and 3 are legal, by checking whether the "from" dates from the date column comes before the dates in the "to" date column. Out of 24 columns, at this moment, we pay more attention to columns related to time and location. To that regard, checking the last four columns do not return satisfactory result since about 195 k rows are missing out of 5.58M rows in total. There were no mismatched values in lat_lon column, meaning that all the records are homogenous in floating values.

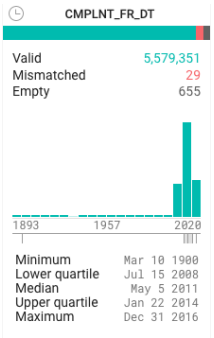

Figure 1: Dataprep result for column 1

3.1.2 NYPD Motor Vehicle Collision data issues. Comparing to the NYPD Compliant Data Historic data set, the NYPD Motor Vehicle Collision was missing values in location-related columns including BOROUGH (column 3), ZIP_CODE (column 4),

LATITUDE (column 5), LONGITUDE (column 6), LOCATION (column 7), and ON_STREET_NAME. Figure 2 describes how the data for column 3 looked on Dataprep. As we attempted to find the correlation between data sets with respect to time and location, it seemed to be problematic to clean this data set at this point since empty cells can't give explanation on how it is related to cells in other data sets. It is worth mentioning that there are several groups of columns that share identical names of different levels. For instance, VEHICLE_TYPE_CODE_1 (column 25) to VEHICLE_TYPE_CODE_4 (column 29) indicate the types of all vehicles involving in a collision. Groups like this tend to start missing values drastically from the secondary column or tertiary column as there was no values recorded in high-level columns in the first place.

The columns for DATE (column 1) and TIME (column 2) appears to be in a decent shape. There are no missing values or mismatched values. In addition, UNIQUE_KEY (column 24) also has neither missing values nor mismatched values and each key is indeed unique.

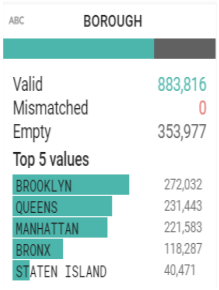

Figure 2: Dataprep result for column 3

3.1.3 Weather data issues. In the case of the weather data, the columns holding the depth of snow accumulation (denoted as SA) was removed since it only contained the default value of 999. Since the Spearman's correlation looks at the monotonic relation between two data sets, having only one value will not help us understand the correlation. For the rows where visibility (Visb), temperature (Temp), and wind speed (Spd) is set to default values, we excluded it as it contains to less information to make accurate analysis. The cells where precipitation (Prcp) or the depth of liquid content of snow that has accumulated on the ground (SDW) is set to the default value of 999.9, we replaced it with 0 since we are going to assume that there has been no precipitation at the time. It is necessary to make this change since the value 999.9 is far greater than any other value in the columns and may cause inaccuracy with determining the correlation. Finally, for the depth of snow and ice on the ground (SD), we have replaced the default value with the last recorded value since we are going to assume that the snow and ice content has remained and there is no means to determine whether the content has decreased (or possibly increased) over time.

3.1.4 Property data issues. When trying to get correlations across different data sets, trying to get the correlation with property data may have been the most difficult. Not only there are places where there are missing values, there were a lot of places where we can't determine if the value is correct. For example, it is hard to determine what is the correct market value (denoted as FULLVAL on the data set) of a property since market values are volatile in nature. Another problem with the data set is that the date is given in years only. This is an issue when correlating with other data sets since we don't know exactly where to aggregate the data. It makes no sense to get correlation of data that starts from January 2010 to June 2014 and data that starts from July 2014 to December 2018 (we are not saying that there is a data like this, it is just an example to show how consistency with date is necessary to a certain extent).

3.1.5 Citi Bike data issues. For the Citi bike data set, the biggest issue was that nearly half of the instances had a different time format from the other half's. In order to create a table on BigQuery, we must keep the time format consistent through the whole date and time column. Another issue was that since 2018, a new column name_localizedValue0 had been added to the monthly trip data. We found this column offering no help for the correlation analysis, therefore, we simply dropped this column separately for the data sets after 2018 and then merged all the data sets into one.

4 RESULTS

4.1 Correlation results

4.1.1 Crime data set. Figure 1 shows the higher-level view of how the frequency of crimes has been reduced year over year. As mentioned in [Cite from NY], a subject of understanding the trend of crime is a compelling subject. Here we did a couple of cross-functional analysis using big data sets on weather, Citi bike, property and collision.

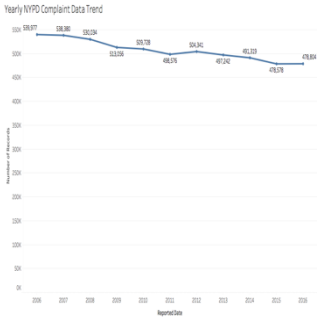

Figure 1: NYPD Complaint Historic total yearly trend

4.1.2 Crime Vs weather. We found a highly correlated result using both Spearman (~0.82) and Pearson (~0.855) between weather pattern and the frequency of reported crime throughout the years (Figure 2 and 3). Intuitively, this is reasonable. Depending on the types of crime, people commit less crimes in warmer weather than colder weather. Hence, we look at the predominant type of crime (Figure 4) and found out that crime type such as Petit Larceny (low-level theft offense), Harassment 2

3

(harassment in the second degree when a person engages in harassing, annoying or alarming to another person in or about public places) and Assault 3 (where a person causes injury to another). All three of them occur mainly in public area.

This result can be supported by William Roberts who wrote an article "The Correlation Between Crime Rates and Weather Patterns in Northern Brooklyn During 2012" [Roberts 2012]. Roberts' hypothesis is that "higher temperature leads to higher rates of property and violent crime". Another research by Anderson has found that higher temperature was "linearly related to assaults" [Anderson 1983].

However, the correlation between wind speed and crime rates are not highly (negatively) correlated. We cannot conclude anything on this regard.

Pearson

	Spd	Visb	Temp	Prcp	SD	SDW
Date	0.082578	0.311675	0.011409	-0.189137	0.060235	0.248818
freq	-0.643294	0.308699	0.805650	0.071478	-0.508989	-0.444608

Spearman

	Spd	Visb	Temp	Prcp	SD	SDW
Date	0.008169	0.367258	0.074024	-0.091055	-0.000566	0.287901
freq	-0.677703	0.377098	0.820085	0.010602	-0.621482	-0.496803

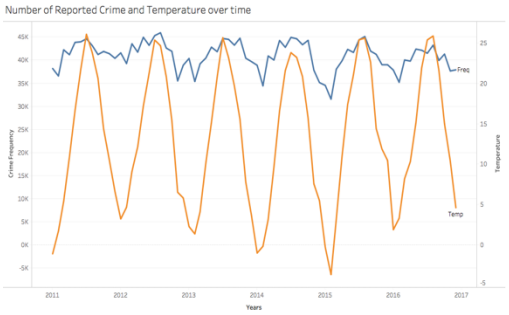

Figure 2: Frequency of Crime Vs Temperature

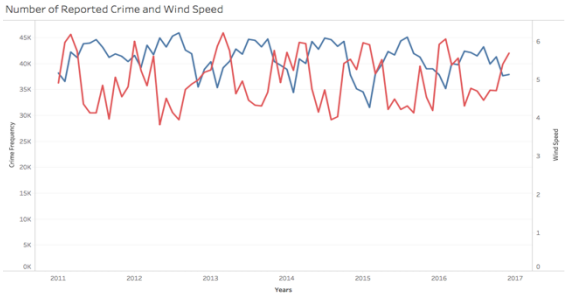

Figure 3: Frequency of Crime Vs Wind Speed

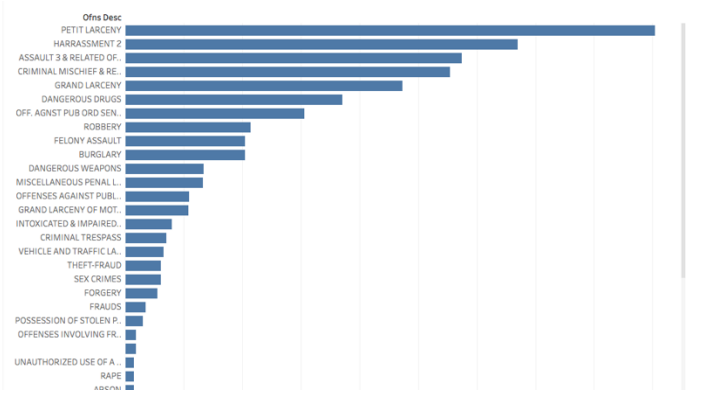

Figure 4: Type of Crime

4.1.3 Crime Vs Collision. On our first attempt, we didn't find the correlation between crime and collision quite notable. However, since both data sets have location column of borough, we took that into consideration and made a more granular table for analysis. It turned out the borough did matter. With a Pearson score of 0.76, a Spearman score of 0.79, and the graph below. we can roughly tell a story that when a crime is committed, an incident of collision is likely to happen.

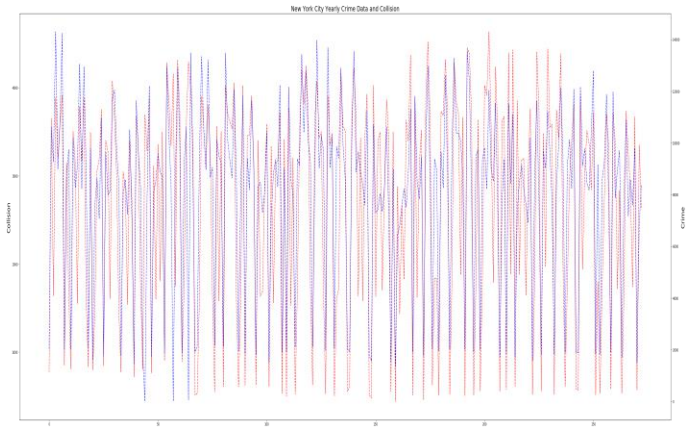

Figure 5: Frequency of Crime Vs Collision with boroughs

4.1.4 Crime Vs Citi Bike. Citi bike sharing statics help us expand our understanding on crime data. Figure 6 clearly shows us the spike of Citi bike usage in New York city. While the crime rate is going down, Citi bike frequency has been rising (Figure 7). When we run correlation, the results are not highly correlated - Spearman (0.571) and Pearson (0.594).

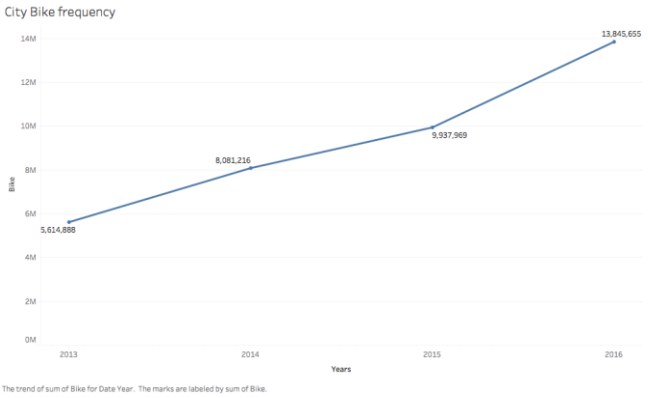

Figure 6: Citi Bike yearly statics

However, we believe that the panel is not rich enough historically. Moreover, Citi bike data starts from 2013 May. Yet, as mentioned before, crime rates have been reduced, unlike the graph below, from 2013 to 2014. Hence, we concluded that if we have richer panel data for both crime and Citi bikes, the correlation result will be expected much higher. Citi bike might be an economic indicator to hint that city's living standard has been improved while the crime rate has been reduced.

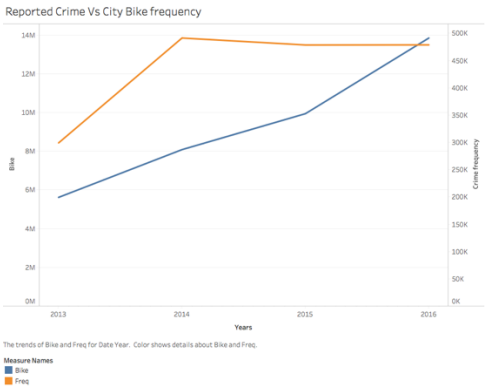

Figure 7: Citi bike and frequency of crime trend

4.1.5 Temperature Vs Citi Bike. We quickly ran the correlation between city bike and temperature. As shown in Figure 8, we can see that two data sets are highly correlated with Spearman (0.7482) and Pearson (0.766).

Since we have previously shown that weather and crime has high correlation and crime and Citi bike has high correlation, it is noteworthy to see how weather and Citi bike has high correlation since it shows associativity to a certain degree.

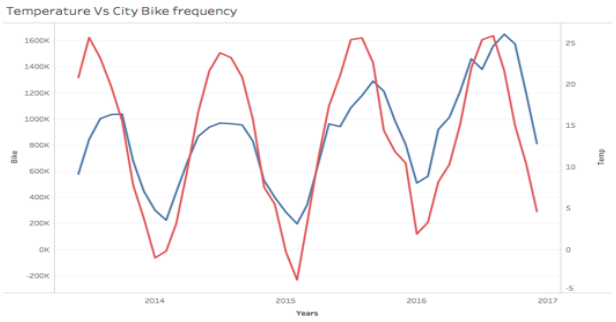

Figure 8: Temperature Vs Citi bike frequency

4.1.6 Crime Vs Property. Property whose price gradually increases gives the very high correlation with the rate of crime. (Figure 9). We also look at the statics between two data sets borough by borough – which again gives us the granular trend.

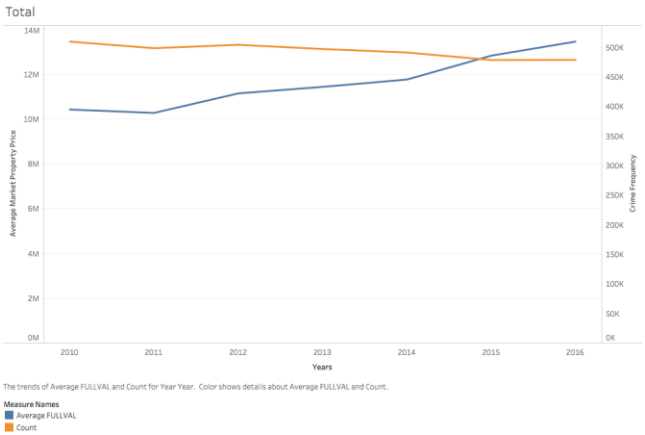

Figure 9: Average Market Property Price Vs Crime statics

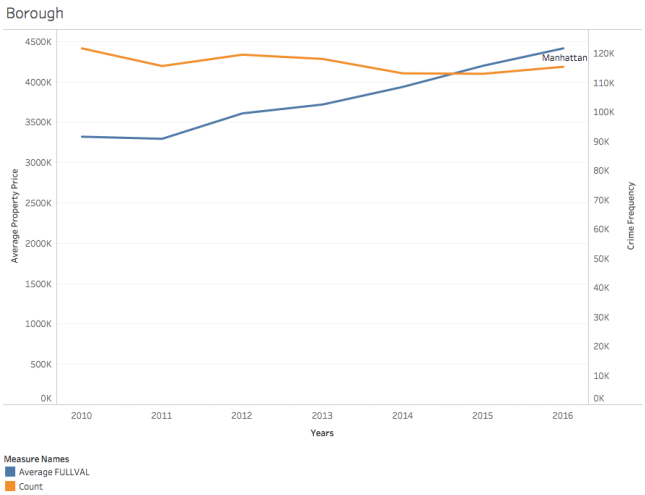

Figure 10: Average Market Property Price Vs Crime statics in Manhattan

5

We have the highest negative correlation with property and crime data (Spearman - 0.85 and Pearson -0.9). We even go further by looking at boroughs by boroughs and here are the satisfying results.

Boroughs	Manhattan	Brooklyn
Spearman	-0.678571	-0.821429
Pearson	-0.635488	-0.892344
Bronx	**Queens**	**Staten Is.**
-0.714286	-0.357143	-0.892857
-0.582812	-0.748927	-0.926933

4.2 Codes

The codes for this project is stored on the link to the GitHub repository below. The file "Dataprep" stores codes used to prepare the data for analysis and the file "DataAnalysis" stores the IPython notebooks used to analyze and get the correlations across data sets. We have several tools to aggregate data outputs since each of the applications have strengths and weaknesses and this has been an excellent opportunity to understand their functionality and our preferences.

https://github.com/PyayAungSan/BigData_Project

5 CONCLUSIONS

After running the tests, we have found several interesting correlations across different data sets. As the project centered on the crime data set, most of the analysis done in the previous section is focused on the results of crime and another data set. The task that took the most time in this project was where we had to prepare the data since we had to determine what are the "correct" values and what will be the problem if we reject rows with "incorrect" values.

Our group believes that this project was a success as we were able to utilize the materials we have learned in class and solve real world problems. Not only we used Hadoop and Spark, we also used IPython Notebook, Tableau, Google BigQuery, and Google Dataprep so we believe that we were able to accomplish another objective of this project, which is to experience and implement new technology.

REFERENCES

[1] Wong, Hongjian, et al. "Crime Rate Inference with fBig Data." *ACM Digital Library*, ACM, 13 Aug. 2016, dl.acm.org/citation.cfm?id=2939736.

[2] "Spearman's Rank Correlation Coefficient." *Wikipedia*, Wikimedia Foundation, 10 Apr. 2018, en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient.

[3] "Spearman's Rank-Order Correlation." *Spearman's Rank-Order Correlation - A Guide to When to Use It, What It Does and What the Assumptions Are.*, statistics.laerd.com/statistical-guides/Spearmans-rank-order-correlation-statistical-guide.php.

[4] Hauke, Jan, and Tomasz Kossowski. "COMPARISON OF VALUES OF PEARSON'S AND SPEARMAN'S CORRELATION COEFFICIENTS ON THE SAME SETS OF DATA."*QUAESTIONES GEOGRAPHICAE*, 2011, www.degruyter.com/downloadpdf/j/quageo.2011.30.issue-2/v10117-011-0021-1/v10117-011-0021-1.pdf+.

[5] Roberts, William. "The Correlation between Crime Rates and Weather Patterns in Northern Brooklyn 2012"

https://www.grin.com/document/305520

[6] Craig, Anderson A., and Anderson C. Dona. "Ambient Temperature and Violent Crime: Tests of the Linear and Curvilinear Hypotheses." *Journal of Personality and Social Psychology*, 10 Feb. 1983, pp. 91–97.

Appendix A: Google Cloud Dataprep

Google Cloud Dataprep is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. Dataprep is serverless and works at any scale. There is no infrastructure to deploy or manage. It appears that Google moved some features of OpenRefine to Dataprep as there is some similarity in transformation functions in both tools. The way Dataprep works is that we create dataflows where recipes can be added. A recipe contains the steps of cleaning the data set and we can see how the recipe work on a sample data set. After finishing editing the recipe, we can simply apply the recipe to the whole data set by running a job.

In addition to preparing the data, Dataprep offers informative visualization and summary on each column. As shown in the body of this paper, Dataprep automatically detects the schemas, data types, possible joins, and anomalies such as missing values, outliers, and duplicates. We can also see how the instances are distributed. In the way, we could save some time writing code and go right to the query or analysis.

Appendix B: Google BigQuery

BigQuery is Google's serverless, highly scalable, low cost enterprise data warehouse designed to make data analysts productive. During this project, BigQuery has been performing superbly on querying dates and locations through millions of rows in tables. Especially for the NYC Yellow Taxi data set which contains over 1 billion rows, BigQuery managed to return the results in just a few seconds. The cost is calculated based on

storage and query size. Since Google has offered a great deal of free usages, we barely spent a dime in this project.

The simplicity is BigQuery's biggest advantage over some other cloud querying tools. Since There is no infrastructure or server to manage on BigQuery, we didn't have to make any optimizations such as configuring the CPU, RAM, or HD etc. Moreover, BigQuery supports Legacy SQL which has some drawbacks comparing to Standard SQL but has been favorable to us as it has some exclusive functions converting data and time to a format we are looking for.

Another way we made use of BigQuery was to connect it with Tableau. Tableau and Google BigQuery allows people to analyze massive amounts of data and get answers fast using an easy-to-use, visual interface. By optimizing the two technologies together, we were able to analyze millions of rows in seconds using visual analysis tools without writing code and without any server-side management.

7